**Principal Components Analysis**

1. Why?
	1. A few ecological factors usually drive a number of patterns and processes simultaneously.
	2. Redundant to try to quantify the effects of everything at once
	3. Want to extract and describe the major independent gradients in the dataset
2. *Unconstrained*:
	1. Not interested in relationships outside the dataset itself (no dependent or independent variables).
	2. Do not inform us of mechanisms, merely describes the major gradients
3. Reduces dimensionality of a data set (*as long as the relationships between variables are linear*)
	1. Originally # of dimensions = # of variables
	2. The more redundancy, the fewer new dimensions
4. New dimensions
	1. Weighted linear combinations (Principal components) that represent gradients of maximum variation
	2. Can interpret the meaning of each component by the variables that define that dimension
5. How it’s done
	1. Projections!
	2. First dimension or axis through the centroid (mean position of all the points in all of the coordinate directions) and maximizes the variance in the longest direction.
	3. Second component perpendicular to the first axis and maximizes the width…
	4. Each component is completely independent and complimentary (non-overlapping venn diagram; orthogonal)
6. Data
	1. Ideally all continuous or all categorical
	2. Each sampling entity must be measured on the same set of variables
	3. More sampling entities that variables (rows > columns)
	4. Examples
		1. Sites – Environmental parameters
		2. Individual – Morphometrics
		3. Species – Behavioral characteristics
7. Assumptions
	1. Not so important if goal is merely descriptive, but if used in statistical inference they are important
	2. Multivariate normality
		1. At a fixed value of all other variables, each variable is normal
		2. No way to test for it
		3. Solutions:
			1. Lots of data
			2. Look at your data!
			3. Test for skewness and normality (Shapiro-Wilk, Kolmogorov-Smirnov) for each variable and the Principal component scores
	3. Independent and random samples
	4. Outliers
		1. I am not a big fan of removing them
		2. Diagnostics:
			1. standard deviation (xi-mean)/sd > 2.5 ?
			2. Look at univariate data and PC scores: stem-and-leaf, boxplot, normal probability plots
	5. Linearity
		1. Solutions: detrended correspondence analysis or detrended PCA
		2. Diagnostic: plot PCA scores and arched or curved configurations are bad news
8. Sample size: N = number of entities sampled, P = number of variables
	1. At least N > P
	2. N = 20+3P
	3. N = 4P
	4. N = 10P (yeah, right…)
9. Deriving the Principal components
	1. Correlation or covariance matrix?
		1. PxP symmetrical matrix
		2. PCA summarizes the structure of this secondary matrix
		3. Covariance gives more weight to variables with larger variances
		4. Correlation matrix
			1. Gives equal weight to all variables
			2. Often preferred and required if the scale or unit of measure is different between variables
			3. Higher correlations = more redundancies = better PCA will work
	2. Eigenvalues (latent roots or characteristic roots)
		1. $\left|R- λI\right|=0$
		2. The extent to which a component accounts for variance in the dataset
		3. *Trace*
			1. Sum of eigenvalues
			2. = P if correlation matrix
			3. = sum of variances if covariance matrix
	3. Eigenvectors (Principal component weights)
		1. $\left[R- λ\_{i}I\right]v\_{i}=0$
		2. Correlation matrix: relative importance of each variable to the PC, proportional to the correlation between the variables and the PC
		3. Covariance: who knows?
10. How many PCs do we keep?
	1. Latent root criterion:
		1. with correlation matrix, do not retain if *λi*< 1
		2. good when 20 ≤ P ≤ 50: too conservative at < 20, too many removed at > 50.
	2. Scree plot: Eigenvalues vs. component number, subjective or difficult to judge
	3. Broken stick criterion
		1. Performed the best
		2. Compared with random expected eigenvalue
		3. $λ\_{i}^{\*}=\sum\_{k=i}^{P}\frac{1}{k}$
		4. If λi ≥ λi\*, keep component *i*
	4. Relative percent variance: subjective
		1. $Φ\_{i}= \frac{λ\_{i}}{P}$
		2. Influenced by the number of variables in the dataset
		3. Also declines with increasing sample size
	5. Significance tests
		1. Resampling nonparametric methods
		2. Jackknife
			1. Remove each sampling entity one at a time and recalculate ***Φ***
			2. Results in N samples of ***Φ***
			3. Calculate the average and standard error for each *Φi*
			4. Mean/SE is a t-statistic with N – 1 degrees of freedom
		3. Bootstrap – little more reliable
			1. Sample the original data set N times with replacement and recalculate ***Φ***
			2. Repeat many times
			3. Calculate the average and standard error for each *Φi*
			4. Mean/SE is a t-statistic with N – 1 degrees of freedom
		4. Randomization test
			1. Scramble the entire dataset by variable and recalculate ***Φ***
			2. Repeat many times
			3. Rank order all *Φi\** and compare with original *Φi*, is it within 95%
	6. Meaningful ecological interpretation?
	7. Keep no more than three.
11. Interpretation
	1. Look at *loadings* (correlations between variables and components):
		1. $s\_{ij}=v\_{ij}\sqrt{λ\_{i}}$
		2. further from zero = variable and component are more related
	2. Rule 1
		1. Variables s < -0.3 or > 0.3 *significant*
		2. Variables s < -0.4 or > 0.4 *more important*
		3. Variables s < -0.5 or > 0.5 *more significant*
		4. N > 50
	3. Rule 2
		1. Variables s < -0.32 or > 0.32 *poor*
		2. Variables s < -0.45 or > 0.45 *fair*
		3. Variables s < -0.55 or > 0.55 *good*
		4. Variables s < -0.63 or > 0.63 *very* *good*
		5. Variables s < -0.71 or > 0.71 *excellent*
	4. Rule 3
		1. As increase N and/or P, the significant value of s decreases
		2. Higher components require larger loadings to be significant
	5. Rule 4: jackknife, bootstrap, or randomization on the loadings
	6. Steps
		1. Highlight all significant loadings
		2. If a variable has no strong loadings, you might want to eliminate it
		3. Assign a name or label to a component that reflects the important variables for that component
	7. Communality
		1. The proportion of a variables variance that is accounted for by the remaining PCs
		2. $c\_{j}= \sum\_{i=1}^{P}s\_{ij}^{2}$
	8. PC scores and plots: useful for additional comparisons and checking assumptions of linearity and independence
	9. Rotation
		1. Because PCs are not always easy to interpret
		2. Redistributes the variance from earlier components to later components
		3. Varimax: stretches out the loadings within columns (components)
		4. Quartimax: stretches out the loadings within rows (variables)
12. Limitations
	1. Difficult to interpret with large community patterns or long environmental gradients
	2. Some variable information may be lost when reducing the number of components
	3. Increase the percent of variance explained in the first component simply by adding more and more nonsense variables